外国語教育研究における第二種信号検出モデル：基本の理解とベイジアンモデリング

草　薙　邦　広
広島大学外国語教育研究センター

1．はじめに

1.1　判断行動に付随する三種類のデータ

外国語に関する学習者の判断行動を観測し，それを定量化することは，外国語教育研究における，もっとも古典的かつ一般的な研究手続きのひとつである[1)]。人間および動物が行なう判断行動の様相は，実にさまざまであり，その種類も多岐にわたるが，本論は，視覚および聴覚提示による言語的刺激の弁別のみにこの用語の使用を限定している。たとえば，文法性判断，語彙性判断，再認課題，または真偽性判断などが本論の対象である。

外国語教育研究において，外国語に関する学習者の判断行動は，学習者がもつ知識や技能を測定するため，あるいは，なんらかの認知的機構を解明するために使用される。前者の場合において，判断行動に由来するデータは，熟達度，文法知識，語彙知識などといった潜在変数の影響を受ける観測変数として，またはそれら理論的概念における操作化の方法としてみなされている。後者の場合において，研究者は，主に刺激，被験者の特性，そして実験手続きなどの操作が，判断行動に及ぼす影響について検証する。これは判断行動を一種の認知的過程とみなし，その認知的過程に関心を寄せるからである。

判断行動に由来するデータは，主として以下の三種類に分けられる（Pleskac & Busemeyer, 2010）。一種類目は，判断結果（response）または課題成績（task performance）である。ほとんどの場合，このデータは，正答確率，被験者に帰属される正答率，誤答率，あるいは刺激に帰属される項目通過率または項目困難度というように，古典的テスト理論の要領によって定量化される。そうでなければ，項目反応理論に属する各モデル，すなわち，ラッシュモデル，二母数モデル，三母数モデルなどをデータにフィットさせることもある。または，後述する信号検出モデル（signal detection theory; theory of signal detection; e.g.,; Green & Swets, 1966; Macmillan & Creelman, 2005; Stanislaw & Todorv, 1999; 国内の外国語教育研究に関する文脈では，Kusanagi, 2014; 草薙・後藤, 2016）[2)] によって，弁別力（sensitivity; e.g., d'）と反応バイアス（response bias; e.g., c, β）といった指標を計算する方法も一般的である。

二種類目のデータは，反応時間（reaction/response/decision time）である。反応時間は，2000年代以降，第二言語習得研究や心理言語学といった関連分野の動向を受けて，外国語教育研究において徐々に導入されるようになった。外国語の学習者と，その学習言語を話す母語話者について比較すると，判断行動の終了に要する時間に大きな差があることがしばしば観測される。このような反応時間における群間の差は，母語および外国語の習得過程や，それらの運用機構における差異に由来するものと解釈される。さらに，解析方法に目を配ると，実験計画法および統計的帰無仮説検定によって，群間における平均値の差のみならず，実験条件毎の平均値について比較したり，指数正規合成分布，対数正規分布，逆ガウス分布などといった連続型確率分布のフィット（e.g., 草薙，2017）によって，その分布の母数について検討されることもある。刺激の種類や

課題条件とグルーピング変数が，従属変数であるところの反応時間に対して統計的交互作用を示すことがある，ということも特筆に値する。刺激におけるある種の言語学的条件が，母語話者の反応時間には影響を及ぼすが，学習者のそれには影響を及ぼさない，といった観測的事実は，特に認知主義を採用する外国語教育研究者や第二言語習得研究者にとって，研究上の重要な足がかりとなる。

三種類目のデータは，主観的評定と呼ばれる種のものである。この種類には，自信評定，確信度，主観的確率，そして主観的測度ないし主観的変数（subjective measure; e.g., Dienes, 2008; Scott & Dienes, 2008; Ziori & Dienes, 2006）といったデータが含まれる。たとえば，ある回答者が，ある刺激への判断結果について，その判断が正答であるという自信があるか，またはそれがないか，といったことを問い，これを定量化したものがこの種にあてはまるデータの典型である。このようなデータは，反応時間と同様に，ある種の認知的過程を解明するためのひとつの視座であると期待される。近年の第二言語習得研究では，自信評定，確信度，そして主観的測度などは，明示的知識および暗示的知識，または意識的知識ないし無意識的知識などといった概念を研究対象とする研究実践において使用されている（e.g., Rebuschat, 2013; Rebuschat & Williams, 2012; Rebuschat, Harmick, Riestenberg, Sachs, & Ziegler, 2015; Tamura, Harada, Kato, Hara, & Kusanagi, 2016）。歴史的に見て，このような研究実践は，人工言語学習研究の方法論（Dienes, 2008; Scott & Dienes, 2008; Ziori & Dienes, 2006）に倣ったものである。

1.2　判断行動に由来するデータの関連

さて，判断行動に由来するこれら三種類のデータが，ほとんどの場合，論理的および確率的な意味において，互いに独立でないということは，非常に重要な観点である（e.g., Pleskac & Busemeyer, 2010; 外国語教育研究の文脈では，草薙・川口，2015; Tamura et al., 2016 など）。たとえば，個人内において，正答確率と反応時間に負の共変関係があることは，速さと正確さの二律背反（speed accuracy tradeoff; e.g, Heitz, 2014; Wickelgren, 1977）と呼ばれる，非常にありふれた現象である。一般的に，正答確率は，速さを優先するような課題操作によって著しく低減する。

また，自信評定と正答確率に，正の共変関係があること，すなわち，「自信があると答えた回答は，正答である確率が高い」といった関係も，広く一般的に見られるものである（e.g., Pleskac & Busemeyer, 2010; Vickers, 1979）。それだけでなく，自信評定と反応時間の間にも，なんらかの共変関係が見られることもある（e.g., Kiani, Corthell, & Shadlen, 2014; Pleskac & Busemeyer, 2010; Vickers & Packer, 1982; Volkman, 1934）。しかしながら，反応時間が長いという条件下において，自信がないと回答する確率が上昇することもあれば，逆に同条件下において，自信があると回答する確率が上昇することもある。この差は，一般に課題条件の操作に起因するとされる。課題条件の操作は，これだけに限らず，上記のさまざまな関係を複雑に仲介することがある。たとえば，速さを犠牲にし，正確さを優先するような，あるいは，より困難な課題条件下において，誤答における平均的な反応時間が，正答における平均的な反応時間よりも，大きな値を示す場合もあれば，逆に，正確さを犠牲にし，速さを優先するような，あるいは，比較的容易な課題条件下において，誤答における平均的な反応時間が，正答における平均的な反応時間よりも，小さな値を示す場合もある（Ratcliff & Rouder, 1998; Pleskac & Busemeyer, 2010）。さらに，主観的評定を被験者にもとめること自体が，正答確率や反応時間データに影響を及ぼすことも知られている（Petrusic & Baranski, 2003）。たとえば，主観的評定を設けると，反応時間と誤答率が増加するこ

ともある。

さて，ここで我々がもつべき関心は，このようなこれら三種類の変数それぞれの複雑な振る舞いや，変数間の関連を，体系的に説明するための手立て，すなわち，端的にいって，数理モデルについてである。冒頭で述べたように，判断行動に由来するデータ，つまり，判断結果，反応時間，そして主観的評定は，現在の外国語教育研究において，かなり一般的なものとなっている。しかしながら，外国語教育に関するこれら各種のデータが，観測において，具体的にどのように振る舞うか，そしてどのような関連をもつかについての関心は，概して高いとはいえない。

一方，数理心理学や認知心理学は，判断行動に伴うデータの振る舞いを説明するための手立てとして，これまでにさまざまな数理モデルを提案してきた。たとえば，信号検出モデルは，刺激の弁別に関する非常に優れた数理モデルのひとつである。しかしながら，信号検出モデルは，反応時間や自信評定についての含意をまったくもたない。次に，Ratcliff による拡散過程モデル（diffusion model; e.g., Ratcliff, 1978, 2002）やそれに類する各種のランダムウォークモデルは，判断結果とその反応時間に優れた近似を与え，速さと正確さの二律背反を説明するモデルである。しかし，残された自信評定との関連をも統一的に説明するモデルではない。

他方で，判断結果と主観的評定の確率的関連を検証するためには，伝統的に，第二種信号検出モデル（type II signal detection model; e.g., Evans & Azzopardi, 2007; Galvin, Podd, Drga, & Whitemore, 2003; Kunimoto, Miller, & Pashler, 2001）やその発展的モデル（e.g., Barrett, Dienes, & Seth, 2013; Maniscalco & Lau, 2012），そしてそれらの周辺的手法（e.g., Clarke, Birdsall, & Tanner, 1959; Nelson, 1984）が一般的に使用されてきた。判断行動における判断結果と主観的評定の確率的関連は，さまざまな学術領域において，一般にメタ認知的弁別性（metacognitive sensitivity; e.g., Fleming & Lau, 2014; Maniscalco & Lau, 2012; Rausch, Müller & Zehetleitner, 2015），メタ認知的正確さ（metacognitive accuracy），第二種弁別力（type II sensitivity），自信－正答率相関（confidence-accuracy correlation）などと呼ばれる。本論は，判断行動に伴う三種類のデータのなかでも，この判断結果と主観的評定の確率的関連，またはメタ認知的弁別性について扱うものである。

1.3　メタ認知的弁別性と外国語教育研究

メタ認知的弁別性に関する研究の歴史は古く，古くは精神物理学（psychophysics）にその起源を見出すことができるとされており（Fleming & Lau, 2014），この用語は，現在，判断行動における判断結果と自信評定の確率的関連を，より抽象的なレベルで表すものとして頻繁に使用されている。本論では，用語としてのメタ認知的弁別性を，この意味のみに限定して使用している。学術領域やそれぞれの研究がもつ関心によって，判断行動における判断結果と自信評定の確率的関連は，意識的知識または明示的知識などと言い換えられることもある。たとえば，ある個人が課題に従事するとして，自信があると評定した試行は，自信がないと評定した試行よりも，より正答である確率が高い，といった確率的依存性，または相関が見られるとき，この個人はメタ認知的弁別性をもつ，または意識的知識ないし明示的知識をもつ，とみなす。逆に，この確率的依存性が見られないとき，当該の個人は，メタ認知的弁別性をもたないであるとか，またはこの確率的独立性が無意識的知識や暗示的知識の証拠である，とみなす。特に人工言語学習研究では，このような確率的独立性を無相関基準（zero-correlation criterion）と呼ぶこともあり，これはメタ知識基準（meta-knowledge criterion）のひとつである（e.g., Dienes, 2008; Scott & Dienes, 2008）。

人工言語習得研究における明示的知識や暗示的知識は，従来の第二言語習得研究における同名

の用語と等しいものであると必ずしも捉えるべきではないが，近年，人工言語学習研究を応用した第二言語習得研究の影響によって，メタ認知的弁別性に類する概念は，国内外の第二言語習得研究者（e.g., Rebuschat, 2013; Rebuschat & Williams, 2012）のみならず，外国語教育研究者の関心を引くものとなってきている。しかしながら，メタ認知的弁別性に関する解析法，特にその数理的な背景については，十分な方法論的関心を集めているとはいえない。そこで本論では，メタ認知的弁別性に関するもっとも主流な解析法のひとつである，第二種信号検出モデルについての基本的な解説を行ない，近年注目を浴びているベイジアンモデリングの要領によってその分析例を示し，外国語教育研究分野における当該の解析法に関する展望について述べることとする。

2．第二種信号検出モデルに関する基本的理解

2.1　第一種課題（Type I Task）と第二種課題（Type II Task）

第一に，第二種信号検出モデルを適用するためには，第一種課題（type I task）と第二種課題（type II task）の両方を揃える実験によるデータが必要である。第一種課題とは，文法性判断課題，語彙性判断課題，再認課題，真偽性判断課題などといった通常の刺激弁別のことであり，与えられた刺激がもつ特性について，被験者の判断を記録したものである。判断結果は，一般に，強制二択（two alternative forced choice, 2AFC）や多肢選択法などによって，離散変数として記録される。以降の節では，2AFC による文法性判断課題を例とする。つまり，ここでの第一種課題の結果（type I response; R_1）は，「文法的」（grammatical）または「非文法的」（ungrammatical）といった値を取る。

第二種課題とは，第一種課題における各試行について，二値の自信評定（自信がある vs. 自信がない），リッカート尺度による多値の自信評定，主観的確率（0%–100%），各種の主観的測度（e.g.,「規則」,「直観」,「親密度」,「当て推量」）などを答える手続きを指す。一般的に，第一種課題における各試行後に，その試行の自信などについて被験者に評定，または選択をさせ，これを記録する。以降の節では，基本的に二値による自信評定を例にする。すなわち，第二種課題の結果（type II response; R_2）は,「自信がある」（high confidence）または「自信がない」（low confidence）といった値を取る。

2.2　第一種の反応テーブル

第一種課題における各試行について，提示される刺激の種類（S）は,「文法的」または「非文法的」であるかのどちらかであり，被験者が下す判断も同様に,「文法的」または「非文法的」であるかのどちらかである。ここで，これら二つの次元に基づいて各試行を分類すると，表 1 のような 2 × 2 のクロス集計表が得られる。この表を，本論では第一種の反応テーブル（type I response table）と呼び，表 1 における type I hit，type I miss，type I false alarm，そして type I correct rejection を，総称して第一種の反応カテゴリー（type I response categories）とする。

表 1　第一種の反応テーブル

	S = “Grammatical”	S = “Ungrammatical”
R_1 = “Grammatical”	Type I Hit	Type I False Alarm
R_1 = “Ungrammatical”	Type I Miss	Type I Correct Rejection

ここで，第一種の反応カテゴリーがもつ確率について考える。文法的な刺激に対して，文法的と判断する確率，HR_1（type I hit ratio）は，条件つき確率[3] を使って，

$$HR_1 = P(R_1 = "grammatical" | S = "grammatical") \quad (1)$$

である。非文法的な刺激に対して，文法的であると判断する確率，すなわち，FAR_1（type I false alarm ratio）は，

$$FAR_1 = P(R_1 = "grammatical" | S = "ungrammatical") \quad (2)$$

となる。

type I hit や type I false alarm は，通常，二律背反の関係にあるため，HR_1 の値が高く，同時に FAR_1 の値が低い場合は，正しく刺激を弁別する能力，すなわち弁別力が総合的に高いといえる。この傾向を数理的に表す指標を与える枠組みのひとつが信号検出モデルである[4]。

2.3　第一種信号検出モデル

第一種課題における信号検出モデルの中でも，もっとも一般的な等分散ガウス信号検出モデル（the equal-variance Gaussian signal detection model）では，上記の HR_1 および FAR_1 の関係より，弁別力指標 d' を計算する（e.g., Macmillan & Creelman, 2005）。まずは，等分散ガウス信号検出モデルの視覚的イメージを図 1 に示す。

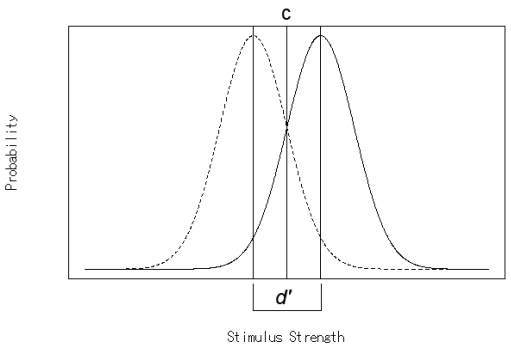

図 1　簡略化した等分散ガウス信号検出モデルの視覚的イメージ

注：横軸は刺激に対する一元的な心理量の強さを表しており，斜線の分布が非文法的な刺激の分布，実線の分布が文法的な刺激の分布である。与えられた刺激に対する心理量が，c を超える場合には文法的と，超えない場合には非文法的と答える。すなわち，実線の内側において，c よりも大きな値を取る範囲の面積が HR，斜線の内側において，c よりも大きな値を取る範囲の面積が FAR となる。等分散の仮定を与えられている二つの正規分布の標準化平均差が d' である。ここでは簡略的にモデルの概観を示すため，二つの分布における確率密度曲線の交点を恣意的に c としているが，c はこの交点とは限らない。

第一種における弁別力指標 d' を，ここでは第一種の弁別力（type I sensitivity）と呼び，d'_1 と記す。d'_1 は，

$$d_1' = z(HR_1) - z(FAR_1) \tag{3}$$

と計算される。ここでの z は，標準正規分布 N (0, 1) における累積分布関数の逆関数（inverse cumulative distribution function）を示す。これは一般的に $\Phi(x)$ と記されることもある。この指標の値は，0 を基準として，正の大きい値を取る場合，より高い弁別力があると解釈される。

また，判断基準ないし反応バイアスの指標である c（criterion）を，第一種課題の場合，第一種の判断基準 c_1 と記す。c_1 は，

$$c_1 = -\frac{1}{2}[z(HR_1) + z(FAR_1)] \tag{4}$$

と計算される。反応バイアスとは，どちらかの判断に対する偏りのことである。たとえば，刺激の種類に独立して，多数の試行に対して「文法的」と答える傾向をもつ被験者は，反応バイアスをもつ。理論上，弁別力は反応バイアスに対して独立した指標であるため，弁別力は，古典的テスト理論の要領によって使用される個人正答率よりも，概して望ましい指標である。反応バイアスとされる指標には，他にも，c' や β などがある。相対基準位置とも呼ばれる c' は，

$$c' = \frac{c}{d'} \tag{5}$$

であり，他方，尤度比にもとづく指標である β は，

$$\beta = e^{cd'} \tag{6}$$

である。ちなみに，ここでの e は，ネイピア数であるから，（6）式は，

$$\ln(\beta) = cd' \tag{7}$$

とも書ける。

2.4　第二種の反応テーブル

第一種課題における反応テーブルを，第二種課題に対して応用することもできる。これが，第二種信号検出モデルの基本的要領である。最初に，第一種課題の結果（R_1）を，正答であるか，誤答であるかの値に変換する。すなわち，type I hit と type I correct rejection を正答，type I miss と type I false alarm を誤答とする。一方，第二種課題の結果（R_2）は，自信がある，自信がないといった値を取る。よって，第一種課題の結果と第二種課題の結果の組み合わせは，表 2 のようにまとめられる。これを，第二種の反応テーブル（type II response table）と呼ぶ。

表２　第二種の反応テーブル

	R_2 = “High Confidence”	R_2 = “Low Confidence”
R_1 = “Correct”	Type II Hit	Type II Miss
R_1 = “Incorrect”	Type II False Alarm	Type II Correct Rejection

上記の type II hit，type II miss，type II false alarm，そして type II correct rejection を，第二種の反応カテゴリー（type II response categories）とする。type II hit は，第一種課題における正答施行のうち，自信があると答えた施行であり，type II miss は，第一種課題における正答施行のうち，自信がないと答えた施行である。また, type II false alarm は, 第一種課題における誤答施行のうち，自信があると答えた施行であり，type II correct rejection は，第一種課題における誤答施行のうち，自信がないと答えた施行である。

先ほどと同様に，これらの確率について考える。HR_2（type II hit ratio）は，

$$HR_2 = P(R_2 = "high\ confidence" | R_1 = "corrrect") \tag{8}$$

であり，これは，

$$HR_2 = P(R_2 = "high\ confidence" | R_1 = Type\ I\ Hit \vee Type\ I\ Correct\ Rejection) \tag{9}$$

とも書ける。一方，FAR_2（Type II false alarm ratio）は，

$$FAR_2 = P(R_2 = "high\ confidence" | R_1 = "incorrrect") \tag{10}$$

または，

$$FAR_2 = P(R_2 = "high\ confidence" | R_1 = Type\ I\ Miss \vee Type\ I\ False\ alarm) \tag{11}$$

となる。HR_2 の値が高く，FAR_2 が低ければ，正答に対して自信があると答え，誤答に対して自信がないと答える傾向が強いといえる。

2.5　第二種信号検出モデル

この第二種の反応テーブルに対して，等分散ガウス信号検出モデルを適用したものが，まさに第二種信号検出モデルである。すなわち，第二種の弁別力（type II sensitivity）d'_2 は，

$$d'_2 = z(HR_2) - z(FAR_2) \tag{12}$$

であり，第二種の判断基準 c_2 は，

$$c_2 = -\frac{1}{2}[z(HR_2) + z(FAR_2)] \tag{13}$$

である。(12) 式で与えられる d'_2 は，もっとも一般的なメタ認知的弁別性の指標となる。すなわち，d'_2 が高ければ，メタ認知的弁別性があるという証拠とみなされる。また，ここでは省略するが，第二種の反応バイアスに関する指標として，c' や β を応用することもできる。

少なくとも理論的には，d'_2 が取る値は，第二種の反応バイアスとは独立したものであるが，これまでの研究により，d'_2 は，技術的に問題のある指標であることがわかっている（e.g., Evans &

Azzopardi, 2007; Galvin et al., 2003）。これは，特に第二種課題のデータにおいて，等分散ガウス信号検出モデルがもつ仮定を満たさない場合があるからである。しかしながら，現状において，第二種信号検出モデルは，もっともスタンダードな手法であり，少なくとも d'_2 は，個人の自信反応率，すなわち，全試行のうち，第二種課題において自信があると答えた試行の比率よりは，メタ認知的弁別性のための優れた指標であるといえる。

3．第二種信号検出モデルの発展系と周辺

3.1　ROC 分析

上記のように，d'_2 は，等分散ガウス信号検出モデルがもつ前提を満たさない指標であると認識されるため，しばしば，ROC 曲線（receiver operating characteristic curve）によるノンパラメトリックな分析が代替的手法として好まれる（Fleming & Lau, 2014）。ROC 曲線を利用した分析法を，ここでは ROC 分析と呼ぶ。ROC 曲線は，横軸に *FAR* を，縦軸に *HR* を置いた空間（ROC 空間）に，バイアスを操作した条件毎に成績をプロットしたものから得られる。よって，ROC 分析をするためには，実験者が報酬を設けたり，刺激の出現率を変化させたりすることによって被験者の反応バイアスを操作し，複数の条件を設ける必要がある。または，多段階評定法といわれる手法を使用することもあり，どちらかといえば，こちらのほうが ROC 分析において簡便な手法である。

多段階評定法では，第二種課題において，自信などについて被験者に段階的な評定をさせる。ここではリッカート尺度の要領によって 5 段階設けるとする（e.g.,「まったく自信がない」,「やや自信がない」,「どちらともいえない」,「やや自信がある」,「非常に自信がある」）。こうして得られた 5 段階の反応について，（a）反応 1 を自信なし，反応 2 以上を自信あり，（b）反応 3 未満を自信なし，反応 3 以上を自信あり，（c）反応 4 未満を自信なし，反応 4 以上を自信あり，（d）反応 5 未満を自信なし，反応 5 を自信あり，というように集計すると，それぞれの *HR* と *FAR* を元にして，ROC 空間に 4 つの点を附置することができる。大雑把にいえば，ROC 空間において，これら 4 つの点を通る曲線が ROC 曲線である。図 2 に，ROC 曲線の例を示す。ROC 空間において，この ROC 曲線下の面積（the area under ROC curve）をもとめると，0 から 1 の値を取るが，この値が 0.5 以上である場合，これはメタ認知的弁別性の証拠となりうる。

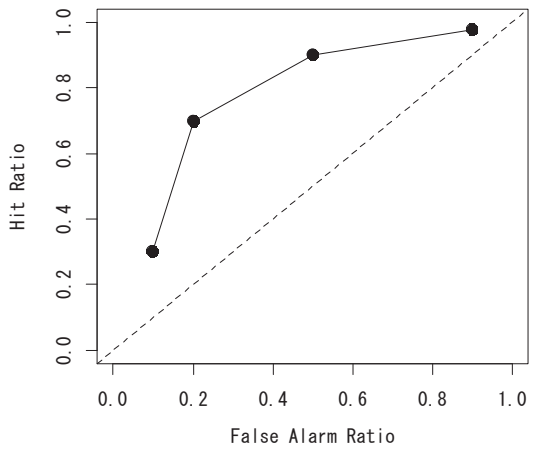

図２　ROC 曲線の例

3.2　近年の発展的モデル

しかしながら，この ROC 曲線下の面積も，第一種の弁別力や反応バイアスから大きな影響を受けることが知られている（Galvin et al., 2003）。近年になって，Maniscalco and Lau（2012）は，meta-*d'* という指標を開発した。数理的な詳細は省略するが，この指標は，第一種課題成績によって調整された第二種の弁別力であるといえ，理論的には第一種課題成績の影響を受けず，頑健であるとされている。

また，d'_1 と meta-*d'* は，同一の単位となるため，それぞれ直接的に値を比較することができる（Fleming & Lau, 2014）。現在，meta-*d'* を使用した研究実践の数は少ないが，今後，このような新しい手法が徐々に市民権を得ていくと推察される。

3.3　周辺的手法

第二種の反応テーブルは，2 × 2 のクロス集計表に他ならないため，より一般的なクロス集計データにおける関連性の分析手法を応用することによっても，メタ認知的弁別性を表すことができるかもしれない。たとえば，第一種課題と第二種課題の成績におけるピアソンの積率相関係数は，Φ 係数と呼ばれる指標でもある。Φ 係数も，メタ認知的弁別性の指標として使用されることがある（Fleming & Lau, 2014）。Φ 係数は，表 3 の場合，

$$\Phi = \frac{AD - BC}{\sqrt{(A + C)(B + D)(C + D)(A + B)}} \tag{14}$$

となる。

表３　クロス集計表として見る第二種の反応テーブル

	R_2 = “High Confidence”	R_2 = “Low Confidence”	Total
R_1 = “Correct”	A	C	A+C
R_1 = “Incorrect”	B	D	B+D
Total	A+B	C+D	A+B+C+D

なお，Φ 係数は，χ^2 検定の統計量を使用し，

$$\Phi^2 = \frac{\chi^2}{n} \tag{15}$$

として計算することもできる。

Goodman-Kruskall の Gamma 係数（*G*）も同様に，メタ認知的弁別性の指標とされることがある（Nelson, 1984）。*G* は，2 × 2 のクロス集計表を対象にする場合，Yule の *Q* と一致する。この場合の *G* は，表 3 を例に取ると，

$$Q = \frac{AD - BC}{AD + BC} \tag{16}$$

となる。同様に，オッズ比，相対リスク，またはテトラコリック相関係数などを分析に応用することができるかもしれない。しかしながら，これらの方法は，反応バイアスの影響を強く受ける

ことが明白である。

４．ベイジアンモデリングによる解析例

ここからは，具体的な数値例を挙げ，分析の様子を概観することとする。標準的な第二種信号検出モデルによる指標は，上記の通りに容易に計算できるが，ベイジアンモデリングの枠組みによって，一般的な信号検出モデルを扱うことができるように（e.g., Lee & Wagenmakers, 2013; 豊田, 2017），第二種信号検出モデルについても，ベイジアンモデリングの枠組みに沿って分析することが可能である。近年の外国語教育研究では，ベイズ統計への関心が飛躍的に高まりつつあるため，ここでは，ベイジアンモデリングを使用した第二種信号検出モデルの分析例を示す。

たとえば，表４のような第二種の反応テーブルが得られたとする。これを，ある被験者が 100 試行の文法性判断課題に従事したものと考える。なお，このデータは，乱数シミュレーションによって生成された仮想的な例であるが，外国語教育研究における過去の研究実績を参照している（e.g., 草薙・川口, 2015; Tamura et al., 2016）。

表４　数値例における第二種の反応テーブル

	R_2 = “High Confidence”	R_2 = “Low Confidence”
R_1 = “Correct”	43	26
R_1 = “Incorrect”	8	23

表４のデータを，頻度主義下の第二種信号検出モデルによって解析すると，$d'_2 = 0.96$，$c_2 = 0.17$ といった値が得られる。一方，ベイジアンモデリングを使用すると，これらの指標における事後分布について詳細に検討することができる。Lee and Wagenmakers（2013）を参考とし，モデルを構築し，このデータの d'_2 および c_2 の事後分布について検討する。このモデルは，具体的に，以下に示される通りである。ここでの θ^h は HR_2，θ^f は FAR_2，h は観測変数であるところの type II hit，f は type II false alarm である。C および I は，それぞれ観測上の正答試行数，誤答試行数を示す。d'_2 および c_2 には，それぞれ Lee and Wagenmakers（2013, p. 158）の事前分布を与えている。

$$d'_{2_i} \sim Gaussian(0, \frac{1}{2}) \tag{17.a}$$

$$c_{2_i} \sim Gaussian(0,2) \tag{17.b}$$

$$\theta_i^h \sim \Phi(\frac{1}{2} d'_{2_i}, -c_{2_i}) \tag{17.c}$$

$$\theta_i^f \sim \Phi(-\frac{1}{2} d'_{2_i}, -c_{2_i}) \tag{17.d}$$

$$h_i \sim Binomial(\theta_i^h, C_i) \tag{17.e}$$

$$f_i \sim Binomial(\theta_i^f, I_i) \tag{17.f}$$

このモデルについて，ギブスサンプリングによるマルコフ連鎖モンテカルロ法（MCMC）[5] を

使用して，各母数の事後分析に近似するサンプルを得る。ここでは，チェイン数を 1，反復回数を 10,000，焼却区間（burn-in interval）を 500 とし，間引き区間（thinning interval）は設けなかった。表 5 は，d'_2 および c_2 に関する MCMC サンプルの概要である。なお，事後分布およびトレース図の様子は，図 3 のとおりであり，Geweke の方法によって MCMC は収束したものと判定した。

最高密度区間（highest density interval; Kruschke, 2011）によって，α = .05 の確信区間[6] を構築したところ，d'_2 については，その下限が 0.39，上限が 1.49 であった。また，c_2 については，その下限が −0.12，上限が 0.44 であった。d'_2 の事後分布における下限が原点よりも高く，事後期待値（expected a posteriori）が 0.93 であることから，この数値例における被験者は，この分析の上で，メタ認知的弁別性を示したと解釈することができる。

表５　個人データにおける MCMC サンプルの要約

	事後期待値	標準偏差	2.5% 点	25% 点	中央値	75% 点	97.5% 点
d'_2	0.93	0.28	0.38	0.73	0.93	1.12	1.49
c_2	0.16	0.14	−0.12	0.06	0.16	0.25	0.44

注：この表ではパーセンタイル法を使用しているが，本文の通り，確信区間の構築には最高密度区間を使用している。

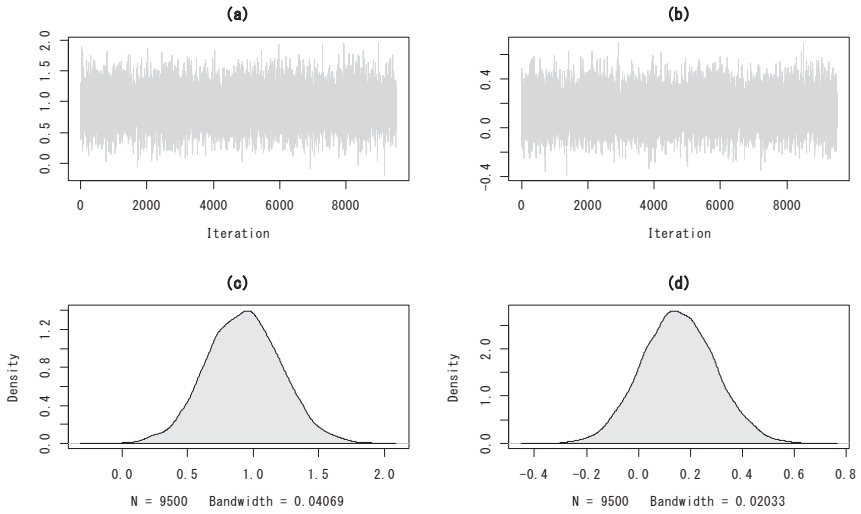

図３　個人データにおけるトレース図と事後分布に近似するサンプルを表すカーネル密度曲線

さて，実際の外国語教育研究では，特定の個人がもつメタ認知的弁別性というよりは，メタ認知的弁性における個人間分散や集団平均が研究者の関心となる。階層的信号検出モデル（e.g., Lee & Wagenmakers, 2013; 豊田，2017）を第二種信号検出モデルに適用することによって，ある集団における d'_2 や c_2 の母平均（μ）および母標準偏差（σ）の事後分布を検討することが可能になる。

Lee and Wagenmakers（2013, p. 162）が示した階層的信号検出モデルを第二種課題に応用すると，以下のように形式化できる。

$$\mu_{d'_2}, \mu_{c_2} \sim Gassian(0, \frac{1}{1000}) \tag{18.a}$$

$$\lambda_{d'_2}, \lambda_{c_2} \sim Gamma(\frac{1}{1000}, \frac{1}{1000}) \tag{18.b}$$

$$d'_{2_i} \sim Gaussian(\mu_{d'_2}, \sqrt{\lambda_{d'_2}}) \tag{18.c}$$

$$c_{2_i} \sim Gaussian(\mu_{c_2}, \sqrt{\lambda_{c_2}}) \tag{18.d}$$

$$\theta_i^h \sim \Phi(\frac{1}{2} d'_{2_i}, -c_{2_i}) \tag{18.e}$$

$$\theta_i^f \sim \Phi(-\frac{1}{2} d'_{2_i}, -c_{2_i}) \tag{18.f}$$

$$h_i \sim Binomial(\theta_i^h, C_i) \tag{18.g}$$

$$f_i \sim Binomial(\theta_i^f, I_i) \tag{18.h}$$

すなわち，前述した個人を対象にしたモデルに加えて，正規分布に従う集団の母平均と，ガンマ分布に従う母標準偏差をもつ正規分布に，それぞれ個人の d'_2 や c_2 が従うということである。ここでは，表 5 の数値例を対象として，上記のモデルについて解析する。ギブスサンプリングを使用し，MCMC によって事後分布に近似するサンプルを得たところ（チェイン数を 1，反復回数を 10,000，焼却区間を 1,000，間引き区間なし），このデータにおける d'_2 および c_2 の母平均と母標準偏差は，表 6 および図 4 のようになった。最高密度区間によって，95% 確信区間を構築すると，d'_2 の母平均における下限は 0.32，その上限は 0.95 であり，d'_2 の母標準偏差における下限は 0.02，その上限は 0.40 であった。また，c_2 の母平均における下限は −0.12，その上限は 0.22 であり，c_2 の母標準偏差における下限は 0.03，その上限は 0.32 であった。

表５　集団を対象にした第二種の反応テーブルにおける数値例（$N = 10$，$K = 48$）

	Type II Hit	Type II False Alarm	Type II Miss	Type II Correct Rejection
被験者 1	18	3	17	10
被験者 2	22	4	9	13
被験者 3	21	6	11	10
被験者 4	29	8	6	5
被験者 5	21	4	3	20
被験者 6	17	8	11	12
被験者 7	23	3	3	19
被験者 8	21	4	4	19
被験者 9	17	4	16	11
被験者 10	22	3	8	15

表6　集団データにおける MCMC サンプルの要約

	事後期待値	標準偏差	2.5% 点	25% 点	中央値	75% 点	97.5% 点
d'_2 の母平均	0.63	0.15	0.34	0.53	0.63	0.73	0.92
c_2 の母平均	0.05	0.09	−0.12	−0.01	0.05	0.1	0.22
d'_2 の母標準偏差	0.16	0.12	0.03	0.07	0.12	0.21	0.48
c_2 の母標準偏差	0.15	0.09	0.03	0.08	0.13	0.2	0.36

それぞれの母数の事後期待値から，この集団に属する被験者の d'_2 は，平均 0.63，標準偏差 0.16 の正規分布に従うと推測することができる。これは，この集団全体の傾向としてのメタ認知的弁別性を認める証拠のひとつとなるだろう。この例が示したように，ベイジアンモデリングの要領によって，第二種信号検出モデルを応用することも可能であり，より柔軟な解析ができるという点において有用である。

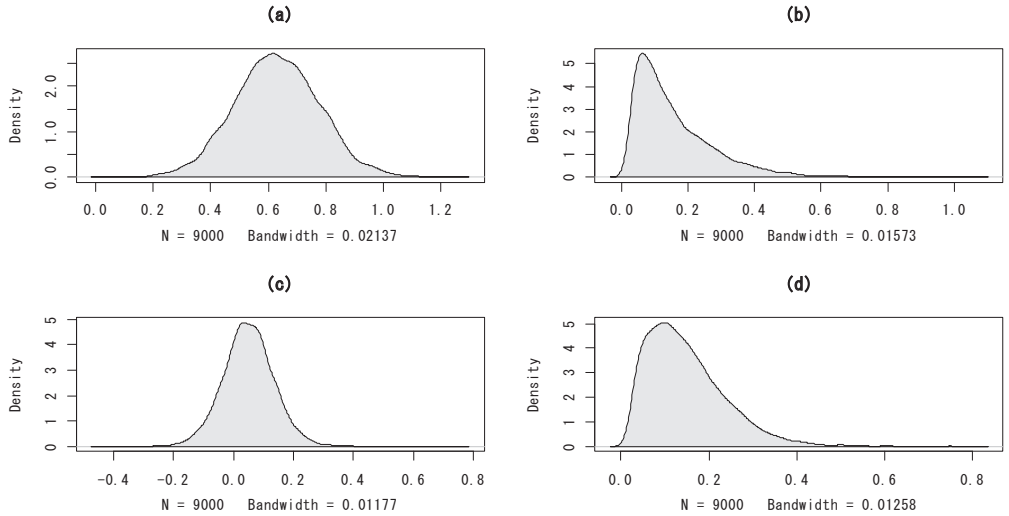

図4　集団データにおける事後分布に近似するサンプルを表すカーネル密度曲線

5．総　括

本論では，判断行動に由来するデータの関連性，その中でも，メタ認知的弁別性とも呼ばれる，判断結果と主観的評定との確率的依存性に焦点を当て，この確率的依存性を分析する手法のひとつである第二種信号検出モデルの基本について概観した。その後，ベイジアンモデリングの要領によって当該のモデルを応用した解析例を示した。これらの概観と解析例が，今後の外国語教育研究における精緻な測定，解析，そして解釈の礎になることを望む。

残念ながら現在のところ，判断結果と主観的評定との確率的依存性，またはメタ認知的弁別性やそれに類する諸概念について，外国語教育研究は豊富な研究実績をもつとはいえない。たとえば，判断結果と主観的評定との確率的依存性に対して，どのような要因が影響を及ぼすのかは，概して不明なままである。影響を及ぼすと合理的に推測できる要因に限っても，(a) 刺激の種類，

(b) 課題条件，(c) 個人についての要因など多種に渡る。また，本論が取り上げなかったデータのひとつである，反応時間を加えた三種類のデータにおける全体的な関連性についても今後の研究の余地がある（cf. Pleskac & Busemeyer, 2010）。さらに，ある個人および集団における判断結果と主観的評定との確率的依存性が，時系列上どのように変化するか，またはある処遇や環境がどのような変化をもたらすか，などといった観点も，外国語教育研究者の関心の外にあるものではない。これらの点に関する今後の実証研究が必要であろう。

注

1）紙面による伝統的なテストのほとんどは，外国語に関する学習者の判断行動を記録したものであるともいえる。

2）このモデルは，一般的に信号検出理論とも呼ばれるが，本論では，これを数理モデルのひとつとみなし，信号検出モデルと一貫して呼んでいる。

3）A と B という事象があるとして，事象 A が起こる確率を *P(A)*，事象 B が起こる確率を *P(B)* と記す。B が起こったときに，または事象 B という条件において，事象 A が起こる確率が条件つき確率，*P(A|B)* である。事象とは，起こりうるすべてのできごとである標本空間の部分集合である。ここでの標本空間は，{A，B，φ} であり，φ は，空事象である。

4）本論は，信号検出モデルを，人間の刺激弁別に関する認知的過程を説明するモデルであると見ているわけではない。認知的過程ではなく，判断行動に由来する観測を説明するための数理モデルのひとつであると考えているため，刺激の特性に対応する一元的な心理量などといったものの物理的実在を強調するわけではない。

5）ここでは，R（R Core Team, 2016）と WinBUGS（Lunn, Thomas, Best, & Spiegelhalter, 2000）を使用している。コードやモデルの詳細は，Lee and Wagenmakers（2013）を参照されたい。

6）Bayesian credible interval のことである。ベイズ信用区間などと訳されることもある。ベイズ統計の枠組みでは，頻度主義とは異なり，観測ではなく母数が分布をなすものと考えるため，95% の確率で，母数はある確信区間の入る，といった解釈を積極的に与えることができる。最高密度区間やパーセンタイル法は，確信区間を構築するための方法である。

参考文献

Barrett, A., Dienes, Z., & Seth, A. K. (2013). Measures of metacognition on signal-detection theoretic models. *Psychological Methods*, *18*, 535–552.

Clarke, F., Birdsall, T., & Tanner, W. (1959). Two types of ROC curves and definition of parameters. *The Journal of the Acoustical Society of America*, *31*, 629–630.

Dienes, Z. (2008). Subjective measures of unconscious knowledge. *Progress in Brain Research*, *168*, 49–64.

Evans, S., & Azzopardi, P. (2007). Evaluation of a “bias-free” measure of awareness. *Spatial Vision*, *20*, 61–77.

Fleming, S. M., & Lau, H. C. (2014). How to measure metacognition. *Frontiers in Human Neuroscience*, *8*, 443.

Galvin, S. J., Podd, J. V., Drga, V., & Whitmore, J. (2003). Type 2 tasks in the theory of signal detectability: Discrimination between correct and incorrect decisions. *Psychonomic Bulletin and Review*, *10*, 843–

876.
Green, D., & Swets, J. (1966). *Signal detection theory and psychophysics*. New York, NY: Wiley.
Heitz, R. P. (2014). The speed-accuracy tradeoff: History, physiology, methodology, and behavior. *Frontiers in Neuroscience*, *8*, 1−19.
Kiani, R., Corthell, L., & Shadlen, M. N. (2014). Choice certainty is informed by both evidence and decision time. *Neuron*, *84*, 1329−1342.
Kruschke, J. K. (2011). Bayesian assessment of null values via parameter estimation and model comparison. *Perspectives on Psychological Science*, *6*, 299−312.
Kunimoto, C., Miller, J., & Pashler, H. (2001). Confidence and accuracy of near-threshold discrimination responses. *Consciousness and Cognition*, *10*, 294−340.
Kusanagi, K. (2014). Speeded effect on accuracy, sensitivity, response bias and reaction time of L2 learners' grammaticality judgments: Using signal detection theory. *JABAET Journal*, *18*, 37−54.
草薙邦広 (2017).「確率分布から見る外国語教育研究データ」『外国語教育メディア学会（LET）関西支部メソドロジー研究部会報告論集』*10*, 1−40.
草薙邦広・川口勇作 (2015).「文法性判断の確信度と明示的および暗示的知識」『中部地区英語教育学会紀要』*44*, 65−72.
草薙邦広・後藤亜希 (2016).「外国語教育研究と信号検出理論」『外国語教育メディア学会（LET）関西支部メソドロジー研究部会報告論集』*8*, 20−36.
Lee, M. D., & Wagenmakers, E. J. (2013). *Bayesian data analysis for cognitive science: A practical course*. New York, NY: Cambridge University Press.
Lunn, D. J., Thomas, A., Best, N., & Spiegelhalter, D. (2000) WinBUGS: a Bayesian modeling framework: concepts, structure, and extensibility. *Statistics and Computing*, *10*, 325−337.
Macmillan, N. A., & Creelman, C. D. (2005). *Detection theory: A users guide* (2nd ed.). New York: Cambridge University Press.
Maniscalco, B., & Lau, H. (2012). A signal detection theoretic approach for estimating metacognitive sensitivity from confidence ratings. *Consciousness and Cognition*, *21*, 422−430.
Nelson, T. (1984). A comparison of current measures of the accuracy of feeling-of-knowing predictions. *Psychological Bulletin*, *95*, 109−133.
Petrusic, W. M., & Baranski, J. V. (2003). Judging confidence influences decision processing in comparative judgments. *Psychonomic Bulletin and Review*, *10*, 177−183.
Pleskac, T. J., & Busemeyer, J. R. (2010). Two-stage dynamic signal detection: a theory of choice, decision time, and confidence. *Psychological Review*, *117*, 864−901.
R Core Team (2016). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.
Ratcliff, R. (1978). Theory of memory retrieval. *Psychological Review*, *85*, 59−108.
Ratcliff, R. (2002). A diffusion model account of reaction time and accuracy in a brightness discrimination task: Fitting real data and failing to fit fake but plausible data. *Psychonomic Bulletin and Review, 9*, 278−291.
Ratcliff, R., & Rouder, J. N. (1998). Modeling response times for two choice decisions. *Psychological Science*, *9*, 347−356.

Rausch, M., Müller, H. J., & Zehetleitner, M. (2015). Metacognitive sensitivity of subjective reports of decisional confidence and visual experience. *Consciousness and Cognition*, *35*, 192−205.

Rebuschat, P. (2013). Measuring implicit and explicit knowledge in second language research. *Language Learning*, *63*, 595−626.

Rebuschat, P., Hamrick, P., Riestenberg, K., Sachs, R., & Ziegler, N. (2015). Triangulating measures of awareness: A Contribution to the debate on learning without awareness. *Studies in Second Language Acquisition*, *37*, 299−334.

Rebuschat, P., & Williams, J. N. (2012). Implicit and explicit knowledge in second language acquisition. *Applied Psycholinguistics*, 33, 829−856.

Scott, R. B., & Dienes, Z. (2008). The conscious, the unconscious, and familiarity. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*, 1264−1288.

Stanislaw, H., & Todorov, N. (1999). Calculation of signal detection theory measures. *Behavior Research Methods, Instruments, and Computers*, *31*, 137−149.

Tamura, Y., & Harada, Y., Kato, D., Hara, K., & Kusanagi, K. (2016). Unconscious but slowly activated grammatical knowledge of Japanese EFL learners: A case of *tough* movement. *Annual Review of English Language Education in Japan*, *27*, 169−184.

豊田秀樹 (2017).『実践ベイズモデリング：解析技法と認知モデル』東京：朝倉書店.

Vickers, D. (1979). *Decision processes in visual perception*. New York, NY: Academic Press.

Vickers, D., & Packer, J. (1982). Effects of alternating set for speed or accuracy on response-time, accuracy and confidence in a unidimensional discrimination task. *Acta Psychologica*, *50*, 179−197.

Volkman, J. (1934). The relation of the time of judgment to the certainty of judgment. *Psychological Bulletin*, *31*, 672−673.

Wickelgren, W. A. (1977). Speed-accuracy tradeoff and information processing dynamics. *Acta Psychologica*, *41*, 67−85.

Ziori, E., & Dienes, Z. (2006). Subjective measures of unconscious knowledge of concepts. *Mind and Society*, *5*, 105−122.

ABSTRACT

The Type II Signal Detection Model in Foreign Language Teaching Research: Some Basics and Bayesian Modeling

Kunihiro KUSANAGI
Institute for Foreign Language Research and Education
Hiroshima University

This paper introduces the type II signal detection model as a means of modeling metacognitive sensitivity of foreign language learners. Foreign language teaching researchers have paid much attention to the conceptual aspects of metacognitive sensitivity, but not to its mathematical formulation and measurement models, while cognitive and mathematical psychologies have voluminous literature on how to model judgment behaviors mathematically. Among mathematical models of judgment behaviors, this paper particularly focuses on the type II signal detection model, one of the most dominant practices of modeling metacognitive sensitivity. The model is nothing but an application of the equal-variance Gaussian signal detection model to type II tasks such as trial-by-trial binary confidence rating during a stimulus discrimination task, which usually corresponds to type I tasks. The model basically calculates two indices, (a) type II *d'* or metacognitive sensitivity and (b) type II biases or metacognitive biases, based on type II response categories below: (a) type II hit, (b) type II false alarm, (c) type II miss, and (d) type II correct rejection. By reviewing such mathematical basics and showing a computation of the model in a Bayesian manner with some numerical examples, this paper claims that the type II signal detection model can be a potentially beneficial tool for foreign language teaching researchers.